

# Exponential Runge-Kutta schemes for inhomogeneous Boltzmann equations with high order of accuracy\*

Qin Li<sup>†</sup> Lorenzo Pareschi<sup>‡</sup>

## Abstract

We consider the development of exponential methods for the robust time discretization of space inhomogeneous Boltzmann equations in stiff regimes. Compared to the space homogeneous case, or more in general to the case of splitting based methods, studied in Dimarco Pareschi [6] a major difficulty is that the local Maxwellian equilibrium state is not constant in a time step and thus needs a proper numerical treatment. We show how to derive asymptotic preserving (AP) schemes of arbitrary order and in particular using the Shu-Osher representation of Runge-Kutta methods we explore the monotonicity properties of such schemes, like strong stability preserving (SSP) and positivity preserving. Several numerical results confirm our analysis.

**Keywords:** Exponential Runge-Kutta methods, stiff equations, Boltzmann equation, fluid limits, asymptotic preserving schemes, strong stability preserving schemes.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>The Boltzmann Equation and its fluid-dynamic limit</b>	<b>3</b>
2.1	Boltzmann Equation . . . . .	3
2.2	Conservations and fluid limit . . . . .	4
<b>3</b>	<b>Exponential Runge-Kutta (ExpRK) methods</b>	<b>5</b>
3.1	Reformulation of the problem and notations . . . . .	5
3.2	Exponential RK schemes with fixed equilibrium function . . . . .	7
3.2.1	The numerical scheme: ExpRK-F . . . . .	7
3.2.2	Choice and evaluation of $\tilde{M}$ . . . . .	8
3.3	Exponential Runge-Kutta schemes with time varying equilibrium function . . . . .	8
3.3.1	The numerical scheme: ExpRK-V . . . . .	9
3.3.2	Computation of $M$ and $\partial_t M$ . . . . .	10

---

\*Research supported by Research Project of National Interest (PRIN 2009) *Advanced numerical methods for kinetic equations and balance laws with source terms*.

<sup>†</sup>Department of Mathematics, University of Wisconsin-Madison, WI, USA

<sup>‡</sup>Department of Mathematics, University of Ferrara, Italy

<b>4</b>	<b>Properties of ExpRK schemes</b>	<b>11</b>
4.1	Positivity and monotonicity properties . . . . .	11
4.2	Contraction and Asymptotic Preservation . . . . .	13
<b>5</b>	<b>Numerical Example</b>	<b>16</b>
5.1	Convergence Rate Test . . . . .	16
5.2	A Sod Problem . . . . .	17
5.3	Mixing Regime . . . . .	19
<b>6</b>	<b>Conclusions and future developments</b>	<b>19</b>
<b>7</b>	<b>Appendix</b>	<b>21</b>
7.1	Positivity of the mass density in ExpRK-V . . . . .	21
7.2	$ P(f) - P(g)  \leq  f - g $ in $d_2$ norm . . . . .	24

## 1 Introduction

The time discretization of kinetic equations in stiff regimes represents a computational challenge in the construction of numerical methods. In fact, in regimes close to the fluid-dynamic limit the collisional scale becomes dominant over the transport of particles and forces the numerical methods to operate with time discretization steps of the order of the Knudsen number. On the other hand the use of implicit integration techniques presents considerable limitations in most applications since the cost required for the inversion of the collisional operator is prohibitive therefore limiting such techniques to simple linear operators.

In recent years there has been a remarkable development of numerical techniques specifically designed for such situations [1, 8, 10, 6, 7, 16, 20]. The basic idea common to these techniques is to avoid the resolution of small time scales by using some a priori knowledge on the asymptotic behavior of the kinetic equation. In particular, we recall among the different possible approaches domain decomposition strategies and hybrid methods at different levels [19, 4, 3, 5, 27].

Asymptotic-preserving schemes have been particularly successful in the construction of unconditionally stable time discretization methods that avoids the inversion of the collision operator. For a nice survey on asymptotic-preserving scheme for various kinds of systems see, for example, the review paper by Shi Jin [15]. In the case of Boltzmann kinetic equations we also refer to the recent review by Pareschi and Russo [21].

In this paper we propose a new class of exponential integrators for the inhomogeneous Boltzmann equation and related kinetic equations which is based on explicit exponential Runge-Kutta methods [14, 17]. More precisely we extend the method recently presented by one of the authors for homogeneous Boltzmann equations [6] to the inhomogeneous case by avoiding splitting techniques. The main feature of the approach here proposed is that it works uniformly with very high-order for a wide range of Knudsen numbers and avoids the solution of nonlinear systems of equations even in stiff regimes. Compared to penalized Implicit-Explicit (IMEX) techniques [8, 7] the main advantage of the class of methods here presented is the capability to easily achieve high order accuracy, asymptotic preservation and monotonicity of the numerical solution.

At variance with the approach presented in Dimarco, Pareschi [6] here we used the Shu-Osher representation of Runge-Kutta methods [26]. This turns out to be essential in order to obtain non splitting schemes with better monotonicity properties (usually referred to as strong stability properties [12]), which permits for example to obtain positivity preserving schemes. In particular we construct methods which are uniformly accurate using two different strategies. The first class of methods is based on the use of a suitable time independent equilibrium state which permits to recover high order accuracy and positivity of the numerical solution. However since the method is based on a constant equilibrium computed at the end time it may suffer of accuracy deterioration in intermediate regimes. The second class of methods is based on computing explicitly the time variation of the Maxwellian state. This permits to obtain schemes with better uniform accuracy but loosing some of the monotonicity property obtained with the first technique.

The rest of the manuscript is organized as follows. In the next section we introduce some preliminary material concerning the Boltzmann equation and its fluid-limit. In Section 3 we derive the novel asymptotic-preserving exponential Runge-Kutta schemes. Two different approaches are presented. The properties of the two approaches are then studied in Section 4. In particular monotonicity properties are investigated. Finally in Section 5 several numerical results for schemes up to third order are presented which show the uniform high order accuracy properties of the present methods. Some theoretical proofs are reported in a separate appendix.

## 2 The Boltzmann Equation and its fluid-dynamic limit

### 2.1 Boltzmann Equation

The Boltzmann equation describes the evolution of the density distribution of rarefied gases. We use  $f(t, x, v)$  to represent the distribution function at time  $t$  on the phase space  $(x, v)$ . The Boltzmann equation is given by

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f), \quad t \geq 0, \quad (x, v) \in \mathbb{R}^d \times \mathbb{R}^d, \quad (2.1)$$

with

$$Q(f, f) = Q^+ - fQ^- = \int_{S^{d-1}} \int_{\mathbb{R}^d} (f'f'_* - ff_*) B(|v - v_*|, \omega) dv_* d\omega. \quad (2.2)$$

Here,  $B$  is the collision kernel,  $\varepsilon > 0$  is the Knudsen number,  $\omega$  is a unit vector, and  $S^{d-1}$  is the unit sphere defined in  $\mathbb{R}^d$  space. We use the shorthands  $f' = f(t, x, v')$  and  $f'_* = f(t, x, v'_*)$ . There are many variations for the collision kernel  $B$ . One simple case is the case of Maxwell molecules when

$$B = B\left(\frac{g \cdot \omega}{|g|}\right),$$

with the relative velocity  $g = v - v_*$ .

The collisional velocities  $v'$  and  $v'_*$  satisfy

$$v' = v - \frac{1}{2}(g - |g|\omega), \quad (2.3a)$$

$$v'_* = v_* + \frac{1}{2}(g - |g|\omega). \quad (2.3b)$$

This deduction is based on momentum and energy conservations

$$\begin{aligned} v + v_* &= v' + v'_*, \\ |v|^2 + |v_*|^2 &= |v'|^2 + |v'_*|^2. \end{aligned}$$

In  $d$ -dimensional space, we define the following macroscopic quantities  $\rho$  is the mass density (here we assume mass is 1, thus number density and mass density have the same value);  $u$  is a  $d$ -dimensional vector that represent the average velocity;  $E$  is the total energy;  $e$  is the specific internal energy;  $T$  is the temperature;  $S$  is the stress tensor; and  $q$  is the heat flux vector, given by

$$\begin{aligned} \rho &= \int f dv, & \rho u &= \int v f dv, \\ E &= \frac{1}{2} \rho u^2 + \rho e = \frac{1}{2} \int |v|^2 f dv, & e &= \frac{d}{2} T = \frac{1}{2\rho} \int f |v - u|^2 dv, \\ S &= \int (v - u) \otimes (v - u) f dv, & q &= \frac{1}{2} \int (v - u) |v - u|^2 f dv. \end{aligned} \quad (2.4)$$

## 2.2 Conservations and fluid limit

Cross section may vary, but the first  $d + 2$  moments of the collision term are always zero. They are obtained by multiplying the collision term with  $\phi = (1, v, \frac{1}{2}|v|^2)^T$  and then integrating with respect to  $v$ , i.e.

$$\begin{aligned} \langle Q \rangle &= \int Q(f) dv = 0, \\ \langle vQ \rangle &= \int vQ(f) dv = 0, \\ \langle \frac{1}{2}v^2Q \rangle &= \int \frac{1}{2}|v|^2 Q(f) dv = 0. \end{aligned} \quad (2.5)$$

Based on these formulas, when taking moments of the Boltzmann equation, one obtains mass, momentum and energy conservation

$$\begin{aligned} \partial_t \rho + \nabla_x \cdot (\rho u) &= \langle Q \rangle = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (S + \rho u^2) &= \frac{1}{\varepsilon} \langle vQ \rangle = 0, \\ \partial_t E + \nabla_x \cdot (Eu + Su + q) &= \frac{1}{\varepsilon} \langle \frac{1}{2}|v|^2 Q \rangle = 0. \end{aligned}$$

For small values of  $\varepsilon$ , the standard Chapman-Enskog expansion around the local Maxwellian

$$M(t, x, v) = \rho(t, x) \left( \frac{1}{2\pi T(t, x)} \right)^{d/2} \exp \left( -\frac{(v - u(t, x))^2}{2T(t, x)} \right), \quad (2.6)$$

shows that at the leading order the moment system yields its Euler limit

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho u) &= 0, \\ \partial_t (\rho u) + \nabla \cdot (\rho u \otimes u + \rho T \mathbb{I}) &= 0, \\ \partial_t E + \nabla \cdot ((E + \rho T)u) &= 0, \end{aligned} \quad (2.7)$$

where  $\mathbb{I}$  is the identity matrix.

### 3 Exponential Runge-Kutta (ExpRK) methods

In this section we would like to extend the Exponential RK method in [6] for the homogeneous Boltzmann equation to the inhomogeneous case (2.1). It has been known for long that time splitting methods degenerate to first order accuracy in the fluid-limit (see [6] and the references therein) so, to achieve high order of accuracy in stiff regimes, time splitting should be avoided.

#### 3.1 Reformulation of the problem and notations

To achieve AP property and robustness in stiff regimes, an implicit method should be adopted. However, due to the complexity and nonlocal property of the collision term  $Q$ , directly inverting it is prohibitively expensive. The Exponential Runge-Kutta method overcomes this difficulty by transforming the equation into the exponential form, and forces the solution to approach to the equilibrium that captures its asymptotic Euler limit as  $\varepsilon$  tends to zero, thus it is an AP scheme. Following the approach in [6], one can define

$$P = Q + \mu f, \quad \mu > 0. \quad (3.1)$$

Let us now consider a nonnegative function  $\tilde{M}$ , hereafter called the *equilibrium function*, and using (2.1) compute

$$\begin{aligned} & \partial_t \left[ (f - \tilde{M}) e^{\mu t / \varepsilon} \right] \\ &= \partial_t (f - \tilde{M}) e^{\mu t / \varepsilon} + (f - \tilde{M}) \frac{\mu}{\varepsilon} e^{\mu t / \varepsilon} \\ &= \left[ \frac{1}{\varepsilon} (Q + \mu f - \mu \tilde{M}) - \partial_t \tilde{M} - v \cdot \nabla_x f \right] e^{\mu t / \varepsilon} \\ &= \left[ \frac{1}{\varepsilon} (P - \mu \tilde{M}) - \partial_t \tilde{M} - v \cdot \nabla_x f \right] e^{\mu t / \varepsilon}. \end{aligned} \quad (3.2)$$

Note that the equation above is equivalent to the original Boltzmann equation (2.1) as long as  $\mu$  is independent on time. In the simplified case of the BGK collision operator  $Q = \mu(M - f)$ , where  $M$  is the local Maxwellian given by (2.6), the problem reformulation just described applies with  $P = \mu M$ . Moreover there is no requirement on the form of  $\tilde{M}$  at all – it can be an arbitrary function. However, to obtain AP property, one has to be careful in picking up its definition, so that the correct asymptotic limit could be captured.

We analyze two different approaches in the following two subsections, and adopt a suitable explicit Runge-Kutta scheme to solve them. For readers' convenience, we firstly give the expression of the Runge-Kutta method used here. Given a large set of ODEs

$$\partial_t y = F(t, y), \quad (3.3)$$

obtained for example using the method of lines from a given PDE, if data  $y^n$  at time step  $t^n$  is known, to compute for the value  $y^{n+1}$  at  $t^{n+1} = t^n + h$ , a classical  $\nu$ -step explicit Runge-Kutta

scheme for equation (3.3) writes

$$\begin{cases} \text{Step } i: & y^{n,(i)} = y^n + h \sum_{j=1}^{i-1} a_{ij} F(t^n + c_j h, y^{n,(j)}), \\ \text{Final step:} & y^{n+1} = y^n + h \sum_{i=1}^{\nu} b_i F(t^n + c_i h, y^{n,(i)}), \end{cases} \quad (3.4)$$

where  $\sum_{j=1}^{i-1} a_{ij} = c_i$ ,  $\sum_i b_i = 1$ , and  $y^{n,(i)}$  stands for the estimate of  $y$  at  $t = t^n + c_i h$ . Different Runge-Kutta method gives different set of coefficients. In the sequel we drop superscript  $n$  for evaluation of  $y$  at sub-stages and use  $y^{(i)} = y^{n,(i)}$ .

Another form of RK method which has proved to be useful in the analysis of the monotonicity properties of Runge-Kutta schemes is the so-called *Shu-Osher representation* [26]. This representation is essential in the study of the positivity properties that will be carried out later

$$\begin{cases} \text{Step } i: & y^{(i)} = \sum_{j=1}^{i-1} \left[ \alpha_{ij} y^{(j)} + h \beta_{ij} F(t^n + c_j h, y^{(j)}) \right], \\ \text{Final step:} & y^{n+1} = \sum_{j=1}^{\nu} \left[ \alpha_{\nu+1,j} y^{(j)} + h \beta_{\nu+1,j} F(t^n + c_j h, y^{(j)}) \right]. \end{cases} \quad (3.5)$$

Let us point out that this latter representation is not unique. Here  $\alpha_{ij}$  are parameters such that  $\sum_{j=1}^{i-1} \alpha_{ij} = 1$ . Without loss of generality, it is natural to set

$$\beta_{ij} = \alpha_{ij} (c_i - c_j), \quad (3.6)$$

for consistency.

**Remark 1.** Expression (3.6) is equivalent with the classical one which says [26]

$$\beta_{ij} = a_{ij} - \sum_{k=j+1}^{i-1} \alpha_{ik} a_{kj}. \quad (3.7)$$

In fact, assume one has  $y^{(j)} = y^n + h \sum_{k=1}^{j-1} a_{jk} F^{(k)}$ ,  $\forall j < i$ , where  $F^{(k)}$  is a shorthand for  $F(t^n + c_k h, y^{(k)})$ , then, by (3.6) one has

$$\begin{aligned} y^{(i)} &= \sum_{j=1}^{i-1} \left[ \alpha_{ij} y^{(j)} + \alpha_{ij} (c_i - c_j) h F^{(j)} \right] \\ &= \sum_{j < i} \left[ \alpha_{ij} \left( y^n + h \sum_{k < j} a_{jk} F^{(k)} \right) + \alpha_{ij} (c_i - c_j) h F^{(j)} \right] \\ &= y^n + h \sum_{j < i} \left( \sum_{k=j+1}^{i-1} \alpha_{ik} a_{kj} + \alpha_{ij} (c_i - c_j) \right) F^{(j)} \end{aligned} \quad (3.8)$$

This clearly requires  $a_{ij} = \alpha_{ij} (c_i - c_j) + \sum \alpha_{ik} a_{kj}$ . Given (3.6), it is  $a_{ij} = \beta_{ij} + \sum \alpha_{ik} a_{kj}$ , which is exactly the classical Shu-Osher representation.

### 3.2 Exponential RK schemes with fixed equilibrium function

Since the choice of the equilibrium function  $\tilde{M}$  in (3.2) is arbitrary, in this subsection, we assume  $\tilde{M}$  as a function independent of time in each time step, i.e.  $\tilde{M}$  is a function given a-priori. Thus (3.2) could be rewritten as

$$\partial_t \left[ (f - \tilde{M}) e^{\mu t/\varepsilon} \right] = \left[ \frac{1}{\varepsilon} (P - \mu \tilde{M}) - v \cdot \nabla_x f \right] e^{\mu t/\varepsilon}. \quad (3.9)$$

Analytically, the equation (3.9) is equivalent to the original inhomogeneous Boltzmann equation as long as  $\tilde{M}$  is a function independent of time and  $\mu$  is a constant. But the associated numerical scheme can preserve asymptotic limit only if  $\tilde{M}$  is chosen in a correct way, as will be clearer later. On the other hand  $\mu$  plays a role in order to guarantee positivity of the numerical solution as will be seen in section.

**Remark 2.** Obviously  $\tilde{M}$  is required not to change in each time step. But for different time steps, we are free to use different functions. This is in fact what we will do, we evolve  $\tilde{M}$  before each time step with a suitable scheme, and then use this computed value function to construct the AP exponential scheme.

#### 3.2.1 The numerical scheme: ExpRK-F

Compared to (3.3),  $y$  turns out to be  $(f - \tilde{M}) e^{\mu t/\varepsilon}$  and the associated evolution function  $F(t, y)$  on the right of (3.3) is  $\left[ \frac{1}{\varepsilon} (P - \mu \tilde{M}) - v \cdot \nabla_x f \right] e^{\mu t/\varepsilon}$ . Thus we have the following scheme

$$\begin{cases} \text{Step } i: & (f^{(i)} - \tilde{M}) e^{c_i \lambda} = (f^n - \tilde{M}) + \sum_{j=1}^{i-1} a_{ij} \frac{h}{\varepsilon} \left[ P^{(j)} - \mu \tilde{M} - \varepsilon v \cdot \nabla_x f^{(j)} \right] e^{c_j \lambda}, \\ \text{Final Step:} & (f^{n+1} - \tilde{M}) e^{\lambda} = (f^n - \tilde{M}) + \sum_{i=1}^{\nu} b_i \frac{h}{\varepsilon} \left[ P^{(i)} - \mu \tilde{M} - \varepsilon v \cdot \nabla_x f^{(i)} \right] e^{c_i \lambda}. \end{cases} \quad (3.10)$$

where we used  $\lambda = \frac{\mu h}{\varepsilon}$ , and  $P^{(j)} = P(f^{(j)})$  for simplicity. Simple algebra gives

- Step  $i$ :

$$f^{(i)} = \left( 1 - e^{-c_i \lambda} - \sum_{j=1}^{i-1} a_{ij} \lambda e^{\lambda(-c_i + c_j)} \right) \tilde{M} + e^{-c_i \lambda} f^n + \sum_{j=1}^{i-1} a_{ij} \lambda e^{\lambda(c_j - c_i)} \left( \frac{P^{(j)}}{\mu} - \frac{\varepsilon}{\mu} v \cdot \nabla_x f^{(j)} \right),$$

- Final Step:

$$f^{n+1} = \left( 1 - e^{-\lambda} - \sum_i b_i \lambda e^{\lambda(-1 + c_i)} \right) \tilde{M} + e^{-\lambda} f^n + \sum_i b_i \lambda e^{\lambda(c_i - 1)} \left( \frac{P^{(i)}}{\mu} - \frac{\varepsilon}{\mu} v \cdot \nabla_x f^{(i)} \right).$$

### 3.2.2 Choice and evaluation of $\tilde{M}$

If it is assumed that

$$0 = c_1 < c_2 < \dots < c_\nu < 1, \quad (3.11)$$

then the same arguments used in [6] shows immediately, that as  $\varepsilon \rightarrow 0$ ,  $\lambda \rightarrow \infty$ , the scheme pushes  $f^{n+1}$  going to  $\tilde{M}$ . So to obtain AP property,  $\tilde{M}$  above should be the Maxwellian at time level  $n+1$  that has the right moments. To get the right moments, the simplest way is to evolve the corresponding macroscopic limit equations, say the Euler equation. We propose solving the Euler equation first to obtain the macroscopic quantities of the Maxwellian for the next time step, and make use of them to define  $\tilde{M}$ . To achieve high order for all regimes, both the macro-solver and micro-solver should be handled by numerical schemes with the same order of accuracy in space and time. The most natural way in time discretization is the explicit Runge-Kutta scheme using the same coefficients as the one for the kinetic equation

$$\left\{ \begin{array}{l} \text{Step } i : \\ \text{Final Step:} \end{array} \right. \quad \left( \begin{array}{c} \rho \\ \rho u \\ E \end{array} \right)^{(i)} = \left( \begin{array}{c} \rho \\ \rho u \\ E \end{array} \right)^n - \Delta t \sum_{j=1}^{i-1} a_{ij} \nabla_x \cdot \left( \begin{array}{c} \rho u \\ \rho u \otimes u + \rho T \\ (E + \rho T) u \end{array} \right)^{(j)}, \quad (3.12)$$

$$\left( \begin{array}{c} \rho \\ \rho u \\ E \end{array} \right)^{n+1} = \left( \begin{array}{c} \rho \\ \rho u \\ E \end{array} \right)^n - \Delta t \sum_{i=1}^{\nu} b_i \nabla_x \cdot \left( \begin{array}{c} \rho u \\ \rho u \otimes u + \rho T \\ (E + \rho T) u \end{array} \right)^{(i)}.$$

**Remark 3.**

- *Note that this method gives us a simple way to couple macro-solver with micro-solver. When  $\varepsilon$  is considerably big, the accuracy of the method is controlled by the micro-solver. And as  $\varepsilon$  vanishes, the method pushes  $f$  going to  $M$ , which is defined by macroscopic quantities computed through the Euler equation while the order of accuracy is given by the macro-solver.*
- *In principle it is possible to adopt other strategies to compute a more accurate time independent equilibrium function in intermediate regions. For example one can use the ES – BGK Maxwellian [9] at time  $n+1$  or one can use the Navier-Stokes equation as the macro-counterpart. Here however we do not explore further in these directions.*
- *The assumption (3.11), although strongly simplifies computations, is in fact not necessary to prove asymptotic preservation. In fact such assumption is independent of the structure of the operator  $P(f, f)$ . We refer to Section 4.2 for more details.*

### 3.3 Exponential Runge-Kutta schemes with time varying equilibrium function

The approach just described has the nice feature of being extremely simple to construct and implement. As we will see in the next section it also possesses several nice features concerning monotonicity. On the other hand it is clear that choosing the limiting equilibrium state in the construction may produce a lack of accuracy in intermediate regimes. To overcome this aspect



here we consider the most natural choice of equilibrium function, namely the local Maxwellian equilibrium state  $\tilde{M} = M$ . The major difficulty in this case is due to the time dependent nature of such equilibrium function.

Now rewrite the equation as

$$\partial_t \left[ (f - M) \exp \left( \frac{\mu t}{\varepsilon} \right) \right] = \left( \frac{P - \mu M}{\varepsilon} - v \cdot \nabla_x f - \partial_t M \right) \exp \left( \frac{\mu t}{\varepsilon} \right), \quad (3.13)$$

and here we define  $\tilde{M}$  has a Gaussian profile that shares the same first  $d + 2$  moments with  $f$ . The moments' equations are governed by

$$\partial_t \int \phi f dv + \int \phi v \cdot \nabla_x f dv = 0, \quad (3.14)$$

with  $\phi = \left[ 1, v, \frac{v^2}{2} \right]^T$ .

### 3.3.1 The numerical scheme: ExpRK-V

The Runge-Kutta method is adopted to solve the system

$$\begin{cases} \partial_t (f - M) e^{\mu t / \varepsilon} &= \frac{1}{\varepsilon} (P - \mu M - \varepsilon v \cdot \nabla_x f - \varepsilon \partial_t M) e^{\mu t / \varepsilon}, \\ \partial_t \int \phi f dv &= - \int \phi v \cdot \nabla_x f dv. \end{cases}$$

Thus we have the following scheme

Step  $i$ :

$$\begin{cases} (f^{(i)} - M^{(i)}) e^{c_i \lambda} &= (f^n - M^n) + \sum_{j=1}^{i-1} a_{ij} \frac{h}{\varepsilon} \left[ P^{(j)} - \mu M^{(j)} - \varepsilon v \cdot \nabla_x f^{(j)} - \varepsilon \partial_t M^{(j)} \right] e^{c_j \lambda}, \\ \int \phi f^{(i)} dv &= \int \phi f^n dv + \sum_{j=1}^{i-1} a_{ij} \left( -h \int \phi v \cdot \nabla_x f^{(j)} dv \right); \end{cases} \quad (3.15a)$$

Final Step:

$$\begin{cases} (f^{n+1} - M^{n+1}) e^\lambda &= (f^n - M^n) + \sum_{i=1}^{\nu} b_i \frac{h}{\varepsilon} \left[ P^{(i)} - \mu M^{(i)} - \varepsilon v \cdot \nabla_x f^{(i)} - \varepsilon \partial_t M^{(i)} \right] e^{c_i \lambda}, \\ \int \phi f^{n+1} dv &= \int \phi f^n dv + \sum_{i=1}^{\nu} b_i \left( -h \int \phi v \cdot \nabla_x f^{(i)} dv \right). \end{cases} \quad (3.15b)$$

The first equation in (3.15a) shows that in each sub-stage  $i$ , to compute for  $f^{(i)}$ , besides the known  $f^{(j)}$  and easily obtained  $M^{(j)}$ , one also needs  $\partial_t M^{(j)}$ ,  $P^{(j)}$ ,  $v \cdot \nabla_x f^{(j)}$  for all  $j < i$ , and  $M^{(i)}$  that is evaluated at the current time sub-stage.

### 3.3.2 Computation of $M$ and $\partial_t M$

Here we show how to compute  $M^{(i)}$  and  $\partial_t M^{(j)}$ .

**Computation of  $M^{(i)}$  :**

solve the second equation of (3.15a), to get evaluation of macroscopic quantities at  $t^n + c_i h$ . Then the Maxwellian  $M^{(i)}$  is given by (2.6).

**Computation of  $\partial_t M^{(j)}$  :**

Write  $\partial_t M$  as

$$\partial_t M = \partial_\rho M \partial_t \rho + \nabla_u M \cdot \partial_t u + \partial_T M \partial_t T, \quad (3.16)$$

and  $\partial_t \rho$ ,  $\partial_t u$  and  $\partial_t T$  can be computed from taking moments of the original equation

$$\begin{aligned} \partial_t \begin{pmatrix} \rho \\ \rho u \\ \frac{d\rho T}{2} + \frac{1}{2}\rho u^2 \end{pmatrix} &= \partial_t \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} M dv = \partial_t \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} f dv \\ &= - \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} v \cdot \nabla_x f dv. \end{aligned} \quad (3.17)$$

To be specific, with data at sub-stage  $(j)$  in  $d$ -dimensional space, one has

$$\partial_t M^{(j)} = \partial_\rho M^{(j)} \partial_t \rho^{(j)} + \nabla_u M^{(j)} \cdot \partial_t u^{(j)} + \partial_T M^{(j)} \partial_t T^{(j)}, \quad (3.18)$$

with

$$\partial_\rho M^{(j)} = \frac{M^{(j)}}{\rho^{(j)}}, \quad \partial_u M^{(j)} = M^{(j)} \frac{v - u^{(j)}}{T^{(j)}}, \quad \partial_T M^{(j)} = M^{(j)} \left[ \frac{(v - u^{(j)})^2}{2(T^{(j)})^2} - \frac{d}{2T^{(j)}} \right], \quad (3.19a)$$

and

$$\partial_t \rho^{(j)} = - \int v \cdot \nabla_x f^{(j)} dv, \quad (3.19b)$$

$$\partial_t u^{(j)} = \frac{1}{\rho^{(j)}} \left( u^{(j)} \int v \cdot \nabla_x f^{(j)} dv - \int v \otimes v \cdot \nabla_x f^{(j)} dv \right), \quad (3.19c)$$

$$\partial_t T^{(j)} = \frac{1}{d\rho^{(j)}} \left( -\frac{2E^{(j)}}{\rho^{(j)}} \partial_t \rho^{(j)} - 2\rho^{(j)} u^{(j)} \partial_t u^{(j)} - \int v^2 v \cdot \nabla_x f^{(j)} dv \right). \quad (3.19d)$$

The  $\partial_t \rho$  and  $\partial_t u$  term in (3.19d) is evaluated by (3.19b) and (3.19c). Clearly, all other macroscopic quantities  $\rho^{(j)}$ ,  $u^{(j)}$  and  $T^{(j)}$  are associated to  $f^{(j)}$ .

## 4 Properties of ExpRK schemes

### 4.1 Positivity and monotonicity properties

Usually positivity, although very important for kinetic equations, is extremely hard to be obtained when using high order schemes. Here we show that thanks to the Shu-Osher representation (3.5) we can follow [11] to prove positivity (and hence SSP property) for the fixed  $\tilde{M}$  method ExpRK-F.

Before proving the theorem we make the following assumption.

**Assumption 1.** *For a given  $f \geq 0$  there exists  $h^* > 0$  such that*

$$f - h v \cdot \nabla_x f \geq 0, \quad \forall 0 < h \leq h^*.$$

The above assumption is the minimal requirement on  $f$  in order to obtain a non negative scheme. Next we can state

**Theorem 1.** *Let us consider an ExpRK-F method defined by (3.10), and  $\beta_{ij} \geq 0$  in (3.6). Then there exist  $h_* > 0$  and  $\mu_* > 0$  such that  $f^{n+1} \geq 0$  provided that  $f^n \geq 0$ ,  $\mu \geq \mu_*$  and  $0 < h \leq h_*$ .*

*Proof.* Using the Shu-Osher representation, one could rewrite the scheme as

$$\begin{cases} \text{Step } i: & (f^{(i)} - \tilde{M})e^{c_i\lambda} = \sum_j e^{c_j\lambda} \left\{ \alpha_{ij}(f^{(j)} - \tilde{M}) + \beta_{ij} \frac{h}{\varepsilon} \left[ P^{(j)} - \mu \tilde{M} - \varepsilon v \cdot \nabla_x f^{(j)} \right] \right\} \\ \text{Final Step:} & (f^{n+1} - M)e^\lambda = \sum_j e^{c_j\lambda} \left\{ \alpha_{\nu+1j}(f^j - \tilde{M}) + \beta_{\nu+1j} \frac{h}{\varepsilon} \left[ P^{(j)} - \mu \tilde{M} - \varepsilon v \cdot \nabla_x f^{(j)} \right] \right\} \end{cases}$$

Simple algebra gives, for  $\forall, i = 1, \dots, \nu, j < i$

$$\begin{aligned} f^{(i)} = & \tilde{M} \left( 1 - \sum_j e^{(c_j - c_i)\lambda} (\alpha_{ij} + \lambda \beta_{ij}) \right) \\ & + \sum_{j=1}^{i-1} \lambda \beta_{ij} e^{(c_j - c_i)\lambda} \frac{P^{(j)}}{\varepsilon} \\ & + \sum_{j=1}^{i-1} \alpha_{ij} e^{(c_j - c_i)\lambda} \left( f^{(j)} - \frac{h \beta_{ij}}{\alpha_{ij}} v \cdot \nabla_x f^{(j)} \right). \end{aligned} \quad (4.1)$$

The same derivation can be also carried out for the final step. If this is a convex combination, then, to have positivity, one only check that each of them is positive

$$\tilde{M} > 0; \quad (4.2a)$$

$$P^{(j)} > 0; \quad (4.2b)$$

$$f^{(j)} - \frac{h \beta_{ij}}{\alpha_{ij}} v \cdot \nabla_x f^{(j)} > 0. \quad (4.2c)$$

Positivity of  $\tilde{M}$  is obvious, and  $P^{(j)}$  is positive if one has big enough  $\mu$

$$\mu \geq \mu_* = \sup |Q^-| \Rightarrow P = Q + \mu f = Q^+ - f Q^- + \mu f > 0.$$

To handle (4.2c), one just need to adopt Assumption 1. It is positive if

$$0 < h \leq h_* = \min_{ij} \left( \frac{\alpha_{ij}}{\beta_{ij}} h^* \right),$$

which guarantees (4.2c).

To check the convexity of (4.1), it should be proved that

$$\sum_j e^{(c_j - c_i)\lambda} (\alpha_{ij} + \lambda\beta_{ij}) \leq 1. \quad (4.3)$$

This can be seen by just taking the derivative with respect to  $\lambda$ . Use  $\Delta_{ij} = c_i - c_j$

$$\begin{aligned} & \frac{d}{d\lambda} \left( \sum_j e^{-\Delta_{ij}\lambda} (\alpha_{ij} + \lambda\beta_{ij}) \right) \\ &= \sum_j e^{-\Delta_{ij}\lambda} (-\Delta_{ij}(\alpha_{ij} + \lambda\beta_{ij}) + \beta_{ij}) \end{aligned} \quad (4.4)$$

$$= \sum_j e^{-\Delta_{ij}\lambda} (-\beta_{ij}\Delta_{ij}\lambda + \beta_{ij} - \alpha_{ij}\Delta_{ij}) < 0 \quad (4.5)$$

In the last step,  $\beta_{ij} = \alpha_{ij}(c_i - c_j)$  is used. Thus the left-hand side of expression (4.3) is monotonically decreasing with respect to  $\lambda \geq 0$  and has a maximum

$$\sum_j \alpha_{ij} = 1$$

at  $\lambda = 0$ . Similarly we can proceed for the final step. This confirms (4.3) and finishes our proof.  $\square$

Since the proof above is based on a convexity argument, we also have monotonicity of the numerical solution or SSP property. Thus the building block of our exponential schemes is naturally given by the optimal SSP schemes which minimize the stability restriction on the time stepping. We refer to [12] for a review on SSP Runge-Kutta schemes.

**Remark 4.**

- Note that the proof above does not rely on the value  $\lambda$  take, i.e. the scheme is positive uniformly in  $\varepsilon$ . For the choice of  $\mu_*$  we refer the reader to the discussion in [6, 10].
- Optimal second and third order SSP explicit Runge-Kutta methods such that  $\beta_{ij} \geq 0$  have been developed in the literature. However the classical third order SSP method by Shu and Osher [26] does not satisfy  $c_j \leq c_i$  for  $j < i$ . Note that standard second order midpoint and third order Heun methods satisfy the assumptions of Theorem 1 (see Table 1.1 page 135 in [13]).
- In [11] it was proved that all four stage, fourth order RK methods with positive CFL coefficient  $h_*$  must have at least one negative  $\beta_{ij}$ . The most popular fourth order method using five stage with nonnegative  $\beta_{ij}$  has been developed in [24]. In [24] the authors also proved that any method of order greater than four will have negative  $\beta_{ij}$ .

- *Positivity of ExpRK-V schemes is much more difficult to achieve because of the involvement of the  $\partial_t M$  term. However, we can prove:*

- (i)  $\rho$  is positive;
- (ii) the negative part of  $T$  is  $O(h\varepsilon)$ .

We leave both the proofs of the above results to the appendix.

## 4.2 Contraction and Asymptotic Preservation

In this section, it will be presented that the new exponential Runge-Kutta schemes preserve the asymptotic limit of the Boltzmann equation. The proof is done by following the proof of contraction in [6].

If one check the formula (3.10) and (3.15b), it seems clear that under assumptions (3.11) the big  $\lambda$  on the shoulder of exponential will push the distance between  $f$  and the Maxwellian function going to zero. But sometimes the Runge-Kutta method may have tough coefficients, say  $c_\nu = 1$ . When this happens, the argument cannot be carried through. However, one could still prove AP property using the particular structure of the collision operator following the framework below.

We need to make use of the following assumption.

**Assumption 2.** *There is a constant  $C$  big enough, such that  $|P(f, f) - P(g, g)| < C|f - g|$  where  $|\cdot|$  denotes a proper metric.*

Part of the proof for the metric  $d_2$  defined in  $P_s(\mathbb{R}^d)$  space (see [28]) can be found in the appendix.

Under this assumption, considering  $P(M, M) = Q(M, M) + \mu M = \mu M$ , one has

$$|P(f, f) - \mu M| < C|f - M|. \quad (4.6)$$

The derivation and the proof for both approaches being AP will be presented below. We first show that ExpRK-F is AP for any given explicit Runge-Kutta scheme.

For AP property, one needs to show that as  $\varepsilon \rightarrow 0$ , the scheme gives correct Euler limit. To do this, basically one needs to prove that  $f$  goes to the Maxwellian function whose macroscopic quantities solve the Euler equation (2.7).

Let us define

$$\begin{aligned} d_i &= |f^{(i)} - \tilde{M}|, & D_i &= |v \cdot \nabla_x f^{(i)}|, & d_0 &= |f^n - \tilde{M}|, & \vec{e} &= [1, 1, \dots, 1]^T, \\ \vec{d} &= [d_1, d_2, \dots, d_\nu], & \vec{D} &= [D_1, D_2, \dots, D_\nu]^T. \end{aligned} \quad (4.7)$$

Moreover  $\mathbb{A}$  is a lower-triangular matrix and  $\mathbb{E}$  is a diagonal matrix given by

$$\mathbb{A}_{ij} = \frac{\lambda}{\mu} a_{ij} e^{(c_j - c_i)\lambda}, \quad \mathbb{E} = \text{diag}\{e^{-c_1\lambda}, e^{-c_2\lambda}, \dots, e^{-c_\nu\lambda}\}.$$

**Lemma 1.** *Based on the definitions above, for ExpRK-F one has*

$$\vec{d} \leq d_0 (\mathbb{I} - C\mathbb{A})^{-1} \cdot \mathbb{E} \cdot \vec{e} + \varepsilon (\mathbb{I} - C\mathbb{A})^{-1} \cdot \mathbb{A} \cdot \vec{D}$$

*Proof.* It is just direct derivation from (3.10)

$$(f^{(i)} - \tilde{M})e^{c_i\lambda} = (f^n - \tilde{M}) + \sum_{j=1}^{i-1} a_{ij} \frac{\lambda}{\mu} e^{c_j\lambda} (P^{(j)} - \mu\tilde{M} - \varepsilon v \cdot \nabla_x f^{(j)}) \quad (4.8)$$

By taking the norm, adopting the triangle inequality, and make use of the assumption that  $|P(f) - \mu\tilde{M}| \leq C|f - \tilde{M}|$ , one gets

$$|f^{(i)} - \tilde{M}| \leq |f^n - \tilde{M}| e^{-c_i\lambda} + \sum_j a_{ij} \frac{\lambda}{\mu} e^{(c_j - c_i)\lambda} \left( C|f^{(j)} - \tilde{M}| + \varepsilon |v \cdot \nabla_x f^{(j)}| \right) \quad (4.9)$$

Written in the matrix form, it becomes

$$\begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_\nu \end{pmatrix} \leq \mathbb{E} \begin{pmatrix} d_0 \\ d_0 \\ \vdots \\ d_0 \end{pmatrix} + C\mathbb{A} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_\nu \end{pmatrix} + \varepsilon\mathbb{A} \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_\nu \end{pmatrix}$$

Thus

$$\vec{d} \leq d_0 \mathbb{E} \cdot \vec{e} + C\mathbb{A} \cdot \vec{d} + \varepsilon\mathbb{A} \cdot \vec{D} \quad (4.10)$$

$$\vec{d} \leq d_0 (\mathbb{I} - C\mathbb{A})^{-1} \cdot \mathbb{E} \cdot \vec{e} + \varepsilon (\mathbb{I} - C\mathbb{A})^{-1} \cdot \mathbb{A} \cdot \vec{D} \quad (4.11)$$

which completes the proof.  $\square$

**Lemma 2.** Define

$$R_1(\lambda) = e^{-\lambda} \left( 1 + \frac{C\lambda}{\mu} \vec{b} \cdot \mathbb{E}^{-1} (\mathbb{I} - C\mathbb{A})^{-1} \mathbb{E} \cdot \vec{e} \right) \quad (4.12)$$

$$\vec{R}_2(\lambda) = \frac{\varepsilon\lambda}{\mu} e^{-\lambda} \vec{b} \cdot \mathbb{E}^{-1} \cdot (\mathbb{I} - C\mathbb{A})^{-1} \cdot (\mathbb{I} + C\mathbb{A}) \quad (4.13)$$

then for scheme (3.10) we have

$$|f^{n+1} - \tilde{M}| \leq |f^n - \tilde{M}| R_1(\lambda) + \vec{R}_2 \cdot \vec{D} \quad (4.14)$$

*Proof.* It is just a simple derivation. Define

$$k_i = \frac{h}{\varepsilon} (P^{(i)} - \mu\tilde{M} - \varepsilon v \cdot \nabla_x f^{(i)}) e^{c_i\lambda}. \quad (4.15)$$

Evidently, the previous lemma leads to

$$|\vec{k}| \leq \frac{\lambda}{\mu} \mathbb{E}^{-1} \cdot (C\vec{d} + \varepsilon\vec{D}). \quad (4.16)$$

Back to (3.10), one has

$$\left(f^{n+1} - \tilde{M}\right) = \left(f^n - \tilde{M}\right) e^{-\lambda} + \sum_{s=1}^{\nu} b_i k_i e^{-\lambda}, \quad (4.17)$$

which implies

$$\left|f^{n+1} - \tilde{M}\right| \leq d_0 e^{-\lambda} + \frac{\lambda}{\mu} e^{-\lambda} \vec{b}^T \cdot \mathbb{E}^{-1} \cdot \left(C \vec{d} + \varepsilon \vec{D}\right) \quad (4.18a)$$

$$\leq e^{-\lambda} \left( d_0 + \frac{\lambda}{\mu} \vec{b} \cdot \mathbb{E}^{-1} \cdot \left( C (\mathbb{I} - C\mathbb{A})^{-1} \cdot \left( d_0 \mathbb{E} \cdot \vec{e} + \varepsilon \mathbb{A} \cdot \vec{D} \right) + \varepsilon \vec{D} \right) \right) \quad (4.18b)$$

$$\leq d_0 e^{-\lambda} \left( 1 + \frac{C\lambda}{\mu} \vec{b} \cdot \mathbb{E}^{-1} \cdot (\mathbb{I} - C\mathbb{A})^{-1} \cdot \mathbb{E} \cdot \vec{e} \right) \quad (4.18c)$$

$$+ \frac{\varepsilon\lambda}{\mu} e^{-\lambda} \vec{b} \cdot \mathbb{E}^{-1} \cdot (\mathbb{I} - C\mathbb{A})^{-1} \cdot (\mathbb{I} + C\mathbb{A}) \cdot \vec{D}. \quad (4.18d)$$

Here  $\vec{b} = [b_1, b_2, \dots, b_\nu]$  is a row vector. The result (4.11) is also used. Plug in the definition of  $R_1$  and  $R_2$ , one gets

$$\left|f^{n+1} - \tilde{M}\right| \leq \left|f^n - \tilde{M}\right| R_1(\lambda) + \vec{R}_2(\lambda) \cdot \vec{D}. \quad (4.19)$$

□

The two lemmas above gives us the estimation of the convergence rate towards the Maxwellian. The smaller  $R_1$  is, the faster the function converges.  $R_2$  represents the drift from the transportation, and is expected to be small in the limit. Also, the matrix  $\mathbb{A}$  is usually a lower triangular matrix, and a strict lower triangular matrix for explicit Runge-Kutta, thus it is a nilpotent.

**Theorem 2.** *The method ExpRK-F defined by (3.10) is AP for general explicit Runge-Kutta method with  $0 \leq c_1 \leq c_2 \leq \dots \leq c_\nu < 1$ .*

*Proof.* Obviously if  $R_1(\lambda) = O(\varepsilon)$  and  $R_2(\lambda) = O(\varepsilon)$  for  $\varepsilon$  small enough, the theorem holds. In fact, for explicit Runge-Kutta method,  $\mathbb{A}$  is a strict lower triangular matrix, and thus a nilpotent, then one has the following

$$\mathbb{E}^{-1} (\mathbb{I} - C\mathbb{A})^{-1} \mathbb{E} = \mathbb{E}^{-1} (\mathbb{I} + C\mathbb{A} + C^2\mathbb{A}^2 + \dots + C^{\nu-1}\mathbb{A}^{\nu-1}) \mathbb{E} \quad (4.20a)$$

$$= \mathbb{I} + \mathbb{B} + \mathbb{B}^2 + \dots + \mathbb{B}^{\nu-1} \quad (4.20b)$$

where  $\mathbb{A}^\nu = 0$ , definition  $\mathbb{B} = C\mathbb{E}^{-1}\mathbb{A}\mathbb{E}$  and  $\mathbb{E}^{-1}\mathbb{A}^2\mathbb{E} = \mathbb{E}^{-1}\mathbb{A}\mathbb{E}\mathbb{E}^{-1}\mathbb{A}\mathbb{E}$  are used. According to the definition of  $\mathbb{A}$  and  $\mathbb{E}$ , it can be computed that

$$\mathbb{B}_{ij} = C\mathbb{A}_{ij} e^{c_i\lambda - c_j\lambda} = \frac{C\lambda}{\mu} a_{ij}.$$

Thus  $\mathbb{I} + \sum_k \mathbb{B}^k$  is a matrix such that: the element on the  $k$ th diagonal is of order  $O(\lambda^k)$ . This leads to obvious result

$$R_1(\lambda) = e^{-\lambda} \left( 1 + \frac{C\lambda}{\mu} \vec{b} \cdot \mathbb{E}^{-1} (\mathbb{I} - \mathbb{A})^{-1} \mathbb{E} \cdot \vec{e} \right) = O(e^{-\lambda} \lambda^{\nu-1}) < O(\varepsilon)$$

Similar analysis can be carried to  $R_2(\lambda)$  to show that it vanishes to zero as  $\varepsilon \rightarrow 0$ .

So as  $\varepsilon \rightarrow 0$ ,  $|f^{n+1} - \tilde{M}| \rightarrow 0$ . By definition,  $\tilde{M}$  is defined by macroscopic quantities computed directly from the limit Euler equation, thus the numerical scheme is AP, which finishes the proof.  $\square$

The derivation of the scheme ExpRK-V is essentially the same, and in the end, one still has, in a condense form

$$|f^{n+1} - M^{n+1}| \leq |f^n - M^n| R_1(\lambda) + \vec{R}_2 \cdot \vec{D} \quad (4.21)$$

with  $R_1$ ,  $\vec{R}_2$ ,  $\mathbb{E}$ ,  $\mathbb{A}$  defined in the same way as in (4.7), but  $D_i = |v \cdot \nabla_x f^{(i)} + \partial_t M^{(i)}|$ . Following the same computations, one could prove that this method is AP too, but the proof is omitted for brevity.

**Theorem 3.** *The method ExpRK-V defined by (3.15) is AP for general explicit Runge-Kutta method.*

## 5 Numerical Example

### 5.1 Convergence Rate Test

In this example, we use smooth data to check the convergence rate of both methods. The problem is adopted from [8]: 1 dimensional in  $x$  and 2 dimensional in  $v$ . Initial distribution is given by

$$f(t=0, x, v) = \frac{\rho_0(x)}{2} \left( e^{-\frac{|v-u_1(x)|^2}{T_0(x)}} + e^{-\frac{|v-u_2(x)|^2}{T_0(x)}} \right) \quad (5.1)$$

with

$$\begin{aligned} \rho_0(x) &= \frac{1}{2} (2 + \sin(2\pi x)), \\ u_1(x) &= [0.75, -0.75]^T, \quad u_2(x) = [-0.75, 0.75]^T, \\ T_0(x) &= \frac{1}{20} (5 + 2 \cos(2\pi x)). \end{aligned}$$

Domain is chosen as  $x \in [0, 1]$  and periodic boundary condition on  $x$  is used. Note that the definition of  $\rho_0$ ,  $u_{1/2}$  and  $T_0$  do not represent the number density, average velocity and temperature.

As one can see, the initial data is summation of two Gaussian functions centered at  $u_1$  and  $u_2$  respectively, and is far away from the Maxwellian. To check the convergence rate, we use  $N_x = 128, 256, 512, 1024$  grid points on  $x$  space, and  $N_v = 32$  points on  $v$  space. Time stepping  $\Delta t$  is chosen to satisfy CFL condition with CFL number being 0.5. We measure the  $L_1$  error of  $\rho$  and compute the decay rate through the following formula [29]

$$\text{error}_{\Delta x} = \max_{t=t^n} \frac{\|\rho_{\Delta x}(t) - \rho_{2\Delta x}(t)\|_1}{\|\rho_{2\Delta x}(t)\|_1}, \quad (5.2)$$

with  $\Delta x = \frac{1}{N_x}$ . Theoretically, a  $k$ th order numerical scheme should give  $\text{error}_{\Delta x} < C (\Delta x)^k$  for  $\Delta x$  small enough.



We compute this problem using spectral method [18] in  $v$ , WENO of order 3/5 [25] for  $x$ . For time discretization, we use the second and third order Runge-Kutta from [13], Table 1 page 135. We denote the four schemes under consideration as ExpRK2-F, ExpRK2-V, ExpRK3-F and ExpRK3-V.

We compute the problem using the Maxwellian, and a distribution function away from the Maxwellian given above as initial data, for  $\epsilon = 1, 0.1, 10^{-3}, 10^{-6}$ . Results are shown in Figure 5.1. We also give the convergence rate Table 5.1. One can see that in kinetic regime, when  $\epsilon = 1$ , the two methods are almost the same, but as  $\epsilon$  becomes smaller, in the intermediate regime, for example  $\epsilon = 0.1$  for the second order schemes and  $\epsilon = 10^{-3}$  for second and third order schemes with Maxwellian data, ExpRK-V performs better than ExpRK-F. In the hydrodynamic regime, however, the two methods give similar results again shown by the two pictures for  $\epsilon = 10^{-6}$ . It is remarkable that the third order methods achieve almost order 5 (the maximum achievable by the WENO solver) in many regimes.

Initial Distribution		Maxwellian Initial		Non-Maxwellian Initial	
$N_x$		128 – 256 – 512	256 – 512 – 1024	128 – 256 – 512	256 – 512 – 1024
$\epsilon = 1$	ExpRK2-F	1.91327	1.99502	1.84968	1.98504
	ExpRK2-V	2.41608	2.02347	2.67733	2.05436
	ExpRK3-F	4.99725	4.35014	5.12959	4.76788
	ExpRK3-V	5.02508	4.40379	5.13515	4.79080
$\epsilon = 0.1$	ExpRK2-F	1.98218	1.99539	1.97725	1.99454
	ExpRK2-V	2.41411	2.02293	2.56620	2.05830
	ExpRK3-F	5.07621	2.94707	5.49587	3.00335
	ExpRK3-V	5.02220	4.39651	5.13859	4.79264
$\epsilon = 10^{-3}$	ExpRK2-F	1.23711	1.64976	1.43331	1.73501
	ExpRK2-V	2.02344	1.85924	1.47466	1.75496
	ExpRK3-F	2.36140	2.69178	2.55225	2.78275
	ExpRK3-V	3.86882	3.03223	2.59114	2.80353
$\epsilon = 10^{-6}$	ExpRK2-F	2.56137	2.04519	2.56137	2.04519
	ExpRK2-V	2.56137	2.04519	2.56383	2.04859
	ExpRK3-F	5.08829	4.56695	5.08830	4.56699
	ExpRK3-V	5.08830	4.56704	4.91909	3.80638

Table 1: Convergence rate for ExpRK methods with different initial data, in different regimes.

## 5.2 A Sod Problem

This simple example is adopted from [29] to check accuracy and AP of the numerical methods. It is a Riemann problem, and the solution to the associated Euler limit is a Sod problem.

$$\begin{cases} (\rho, u_x, u_y, T) = (1, 0, 0, 1), & \text{if } x < 0; \\ (\rho, u_x, u_y, T) = (1/8, 0, 0, 1/4), & \text{if } x > 0; \end{cases}$$

In Figure 5.2 (left), we show that when  $\epsilon = 0.01$  is comparably big, both the two new method proposed here match with the numerical results given by explicit scheme with dense mesh. Here the reference is given by Forward Euler with  $\Delta x = 1/500$  and  $h = 0.0001$ . In Figure 5.2 (right),

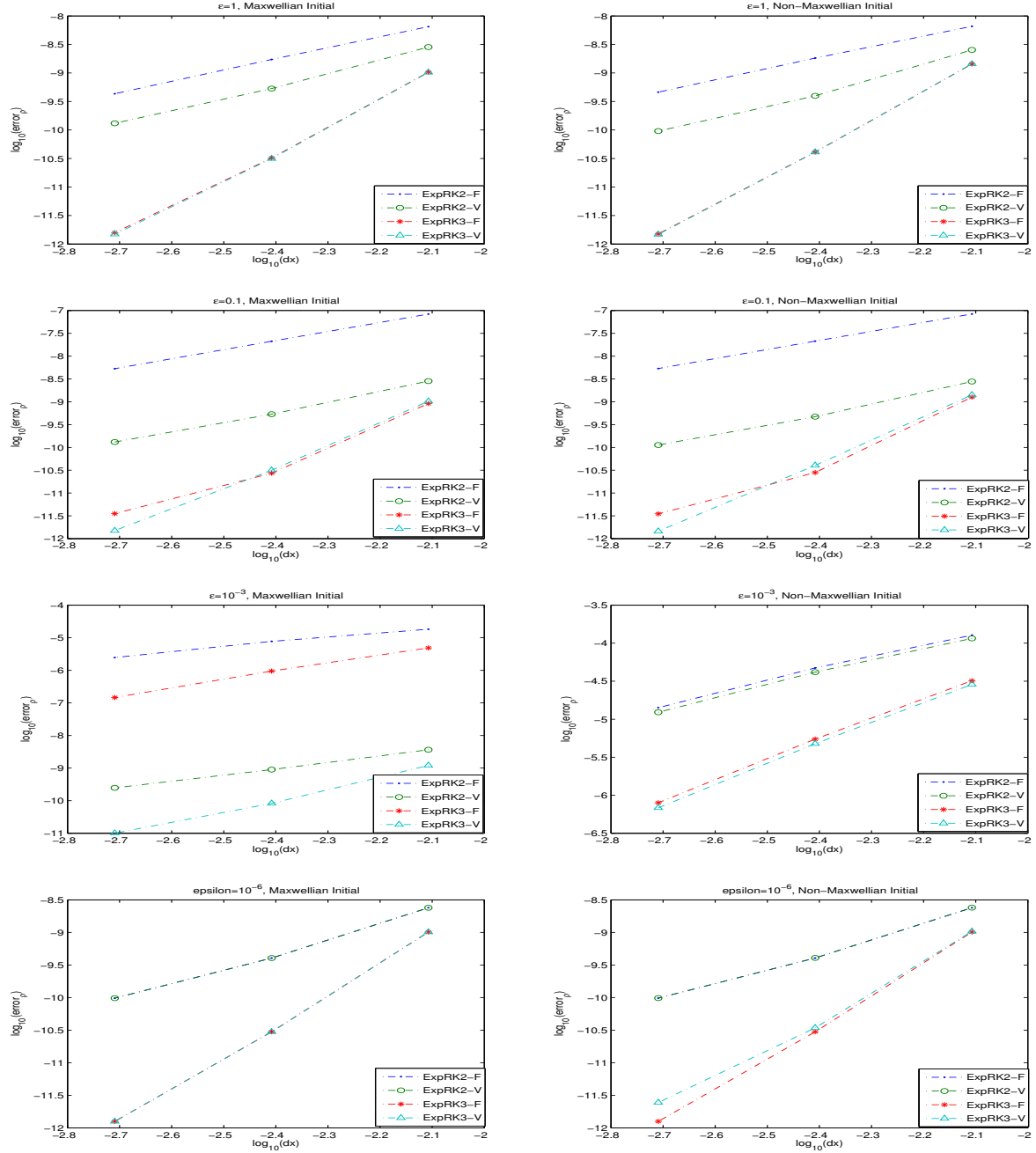


Figure 5.1: Convergence rate test. In each picture, 4 lines are plotted: the lines with dots, circles, stars and triangles on them are given by results of ExpRK2-F, ExpRK2-V, ExpRK3-F and ExpRK3-V respectively. The left column is for Maxwellian initial data, and the right column is for initial data away from Maxwellian (5.1). Each row, from the top to the bottom, shows results of  $\epsilon = 1/0.1/10^{-3}/10^{-6}$  respectively.

AP property is shown: it is clear that for  $\epsilon = 10^{-6}$ , numerical results capture the Euler limit – the Euler limit is computed by kinetic scheme [22]. All plots are given at time  $t = 0.2$ .

### 5.3 Mixing Regime

In this example [29], we show numerical results to a problem with mixing regime. This problem is difficult because  $\epsilon$  vary with respect to space. As what we do in the first example, we take identical data along one space direction, so it is  $1D$  in space but  $2D$  in velocity. An accurate AP scheme should be able to handle all  $\epsilon$  with considerably coarse mesh. Domain is chosen to be  $x \in [-0.5, 0.5]$ , with  $\epsilon$  defined by

$$\epsilon = \begin{cases} \epsilon_0 + 0.5 (\tanh(6 - 20x) + \tanh(6 + 20x)) & x < 0.2; \\ \epsilon_0 & x > 0.2 \end{cases} \quad (5.3)$$

where  $\epsilon_0$  is  $10^{-3}$ . So  $\epsilon$  raise up from  $10^{-3}$  to  $O(1)$ , and suddenly drop back to  $10^{-3}$  as shown in Figure (5.3). Initial data is the give as

$$f(t = 0, x, v) = \frac{\rho_0(x)}{4\pi T_0(x)} \left( e^{-\frac{|v-u_0(x)|^2}{2T_0(x)}} + e^{-\frac{|v+u_0(x)|^2}{2T_0(x)}} \right) \quad (5.4)$$

with

$$\begin{cases} \rho_0(x) = \frac{2+\sin(2\pi x+\pi)}{3}, \\ u_0(x) = \frac{1}{5} \begin{pmatrix} \cos(2\pi x + \pi) \\ 0 \end{pmatrix}, \\ T_0(x) = \frac{3+\cos(2\pi x+\pi)}{4} \end{cases} \quad (5.5)$$

Periodic boundary condition on  $x$  is applied.

We compute the problem using both methods proposed in this paper together with standard explicit Runge-Kutta 2 and 3 in time used as the underline methods in the exponential schemes.

Results are plotted in Figure 5.4. The reference solution is computed with a very fine mesh in time. Both methods give excellent results simply taking a CFL condition of 0.5 whereas explicit methods are forced to operate on a time scale 1000 times smaller. In particular, ExpRK3-V performs well uniformly on  $\epsilon$  by giving a more accurate description of the shock profiles.

## 6 Conclusions and future developments

In this paper we have presented a general way to construct high-order time discretization methods for the Boltzmann equations in stiff regimes which avoid the inversion of the collision operator. The main advantages compared to other methods presented in the literature is the capability to achieve high order uniformly with respect to the small Knudsen number and to originate monotone schemes thanks to the exponential structure of the coefficients. The approach presented here can be extended in principle to several other integro-differential kinetic equations where it is possible to identify a linear operator which preserves the asymptotic behavior of the system. For example in the case of the Landau equation this would involve the computation of the exact flow of the linear part, i.e. a matrix exponential, in the construction of the schemes. We leave this possibility to future research.

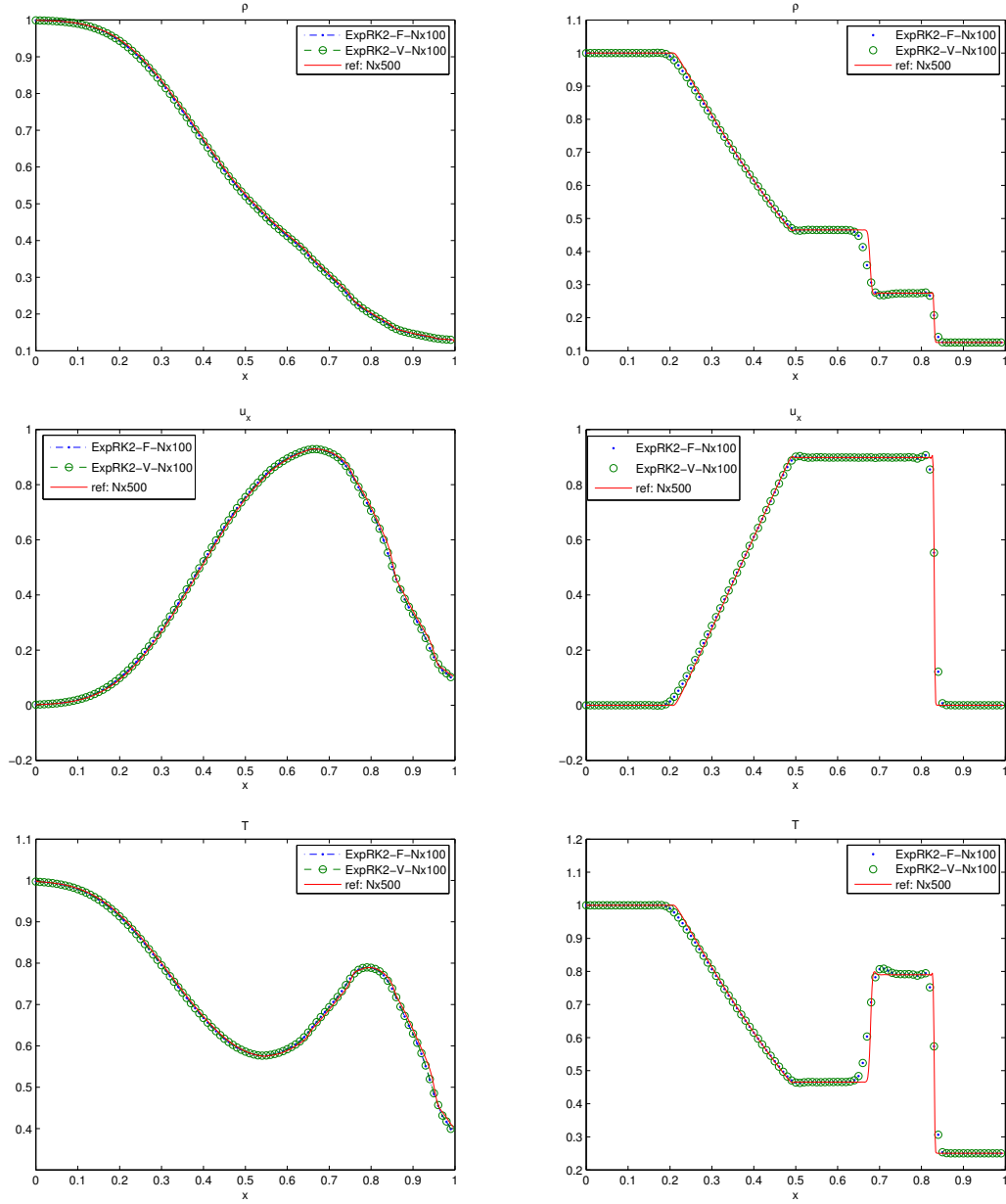


Figure 5.2: Consistency and AP. Left column:  $\varepsilon = 0.01$ . The solid line is given by explicit scheme with dense mesh, while dots and circles are given by ExpRK2-F and ExpRK2-V respectively, both with  $N_x = 100$ .  $h = \Delta x/20$  satisfies the CFL condition with CFL number being 0.5. Right column: For  $\varepsilon = 10^{-6}$ , both methods capture the Euler limit. The solid line is given by the kinetic scheme for the Euler equation, while the dots and circles are given by ExpRK2-F and ExpRK2-V. They perform well in rarefaction, contact line and shock.

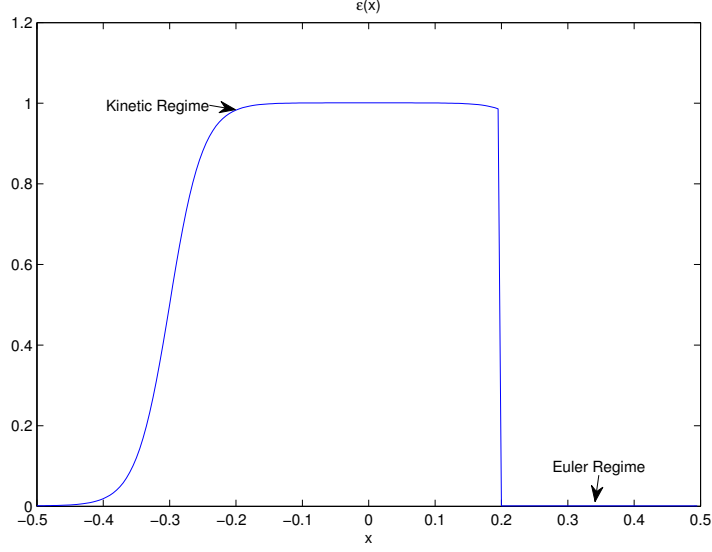


Figure 5.3: Mixing Regime:  $\varepsilon(x)$

## Acknowledgements

The first author would like to thank Dr. G. Dimarco and Prof. S. Jin for stimulating discussions, and Dr. Bokai Yan for providing the code of spectral method for the collision term.

## 7 Appendix

### 7.1 Positivity of the mass density in ExpRK-V

**Theorem 4.** *The method ExpRK-V defined by (3.15) gives positive  $\rho$ , and the negative part of  $T$  is at most of order  $O(h\varepsilon)$ .*

To prove this theorem, we firstly check the following lemma.

**Lemma 3.** *In each sub-stage, the distribution function  $f^{(i)}$  and  $M^{(i)}$  have the same first  $d + 2$  moments.*

*Proof.* We prove this for sub-stage  $i$ . Assume for  $\forall j < i$ , one has

$$\int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} (f^{(j)} - M^{(j)}) dv = 0. \quad (7.1)$$

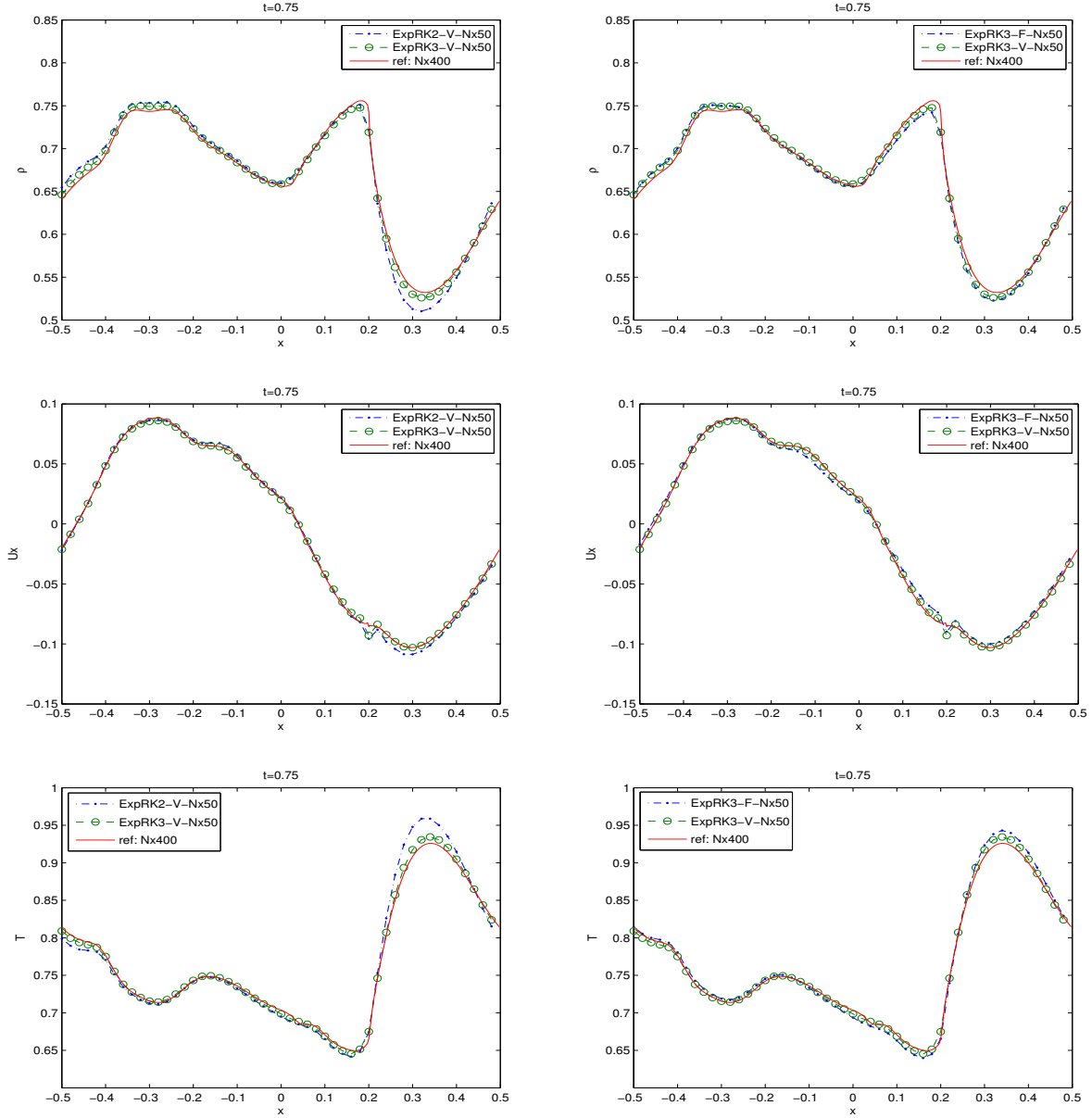


Figure 5.4: The left column shows comparison of RK2 and RK3 using the ExpRK-V. The solid line is the reference solution with a very fine mesh in time and  $\Delta x = 0.005$ , the dash line is given by RK3 and the dotted line is given by RK2, both with  $N_x = 50$  points. The right column compare two methods, both given by RK3, with the reference. The dash line is given by ExpRK-V, and the dotted line is given by ExpRK-F.  $N_x = 50$  for both.  $h$  is chosen to satisfy CFL condition, in our case, the CFL number is chosen to be 0.5.

Then, one could take moments of the first equation in the scheme (3.15a), and gets

$$\int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} (f^{(i)} - M^{(i)}) e^{c_i \lambda} dv = \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} (f^n - M^n) dv \quad (7.2a)$$

$$+ \sum_{j=1}^{i-1} a_{ij} \frac{\lambda}{\mu} e^{c_j \lambda} \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} (P^{(j)} - \mu M^{(j)}) dv \quad (7.2b)$$

$$- \sum_{j=1}^{i-1} a_{ij} \frac{\lambda}{\mu} e^{c_j \lambda} \int \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix} (\varepsilon v \cdot \nabla_x f^{(j)} - \varepsilon \partial_t M^{(j)}) dv \quad (7.2c)$$

(7.2a) is zero for sure, (7.2b) is zero by definition of  $P$  and (7.1). (7.2c) is zero because of the computation from (3.18). Thus it is obvious that  $f^{(i)}$  and  $M^{(i)}$  share the same moments on each stage.  $\square$

With the previous lemma in hand, one could prove Theorem 4.

*Proof.* As in the previous lemma, we only do the proof for sub-stage  $i$ . The final step can be dealt with in the same way. Rewrite the second equation of (3.15a) in Shu-Osher representation

$$\int \phi f^{(i)} dv = \sum_{j=1}^{i-1} \left( \alpha_{ij} \int \phi f^{(j)} dv + \beta_{ij} h \int \phi v \cdot \nabla_x f^{(j)} dv \right) \quad (7.3)$$

This moment equation is the same as the equation on  $\rho$  in the Euler system, and the classical proof for  $\rho$  being positive for the Euler equation can just be adopted [11]. To check the positivity of  $T$ , one just need to make use of the last line of the moment equation, i.e.

$$\begin{aligned} \int \frac{v^2}{2} f^{(i)} dv &= \sum_{j=1}^{i-1} \left( \alpha_{ij} \int \frac{v^2}{2} f^{(j)} dv + \beta_{ij} h \int \frac{v^2}{2} v \cdot \nabla_x f^{(j)} dv \right) \\ &= \sum_{j=1}^{i-1} \left( \alpha_{ij} \int \frac{v^2}{2} f^{(j)} dv + \beta_{ij} h \int \frac{v^2}{2} v \cdot \nabla_x M^{(j)} dv \right) \end{aligned} \quad (7.4a)$$

$$+ h \sum_{j=1}^{i-1} \beta_{ij} \int \frac{v^2}{2} v \cdot \nabla_x (f^{(j)} - M^{(j)}) dv \quad (7.4b)$$

(7.4a) is exactly what one could get when computing for  $E$  in the Euler system: the form of  $M$  closes it up. So the classical method to prove that  $E > \frac{\rho u^2}{2}$  in Runge-Kutta scheme could be used, and the only thing new is from (7.4b). However, as proved in the section about AP, the difference between  $f$  and  $M$  is at most of  $\varepsilon$ , thus (7.4b) is of order  $O(h\varepsilon)$ .  $\square$

## 7.2 $|P(f) - P(g)| \leq |f - g|$ in $d_2$ norm

We adopt the results from [28]. They denote  $P_2$  the collection of distributions  $F$  such that

$$\int_{R^d} |v|^2 dF(v) < \infty$$

A metric  $d_2$  on  $P_2$  is defined by

$$d_2(F, G) = \sup_{\xi} \frac{\hat{f}(\xi) - \hat{g}(\xi)}{|\xi|^2} \quad (7.5)$$

where  $\hat{f}$  is the Fourier transform of  $F$

$$\hat{f}(\xi) = \int e^{-i\xi \cdot v} dF(v)$$

One can transform the Boltzmann equation into its Fourier space and obtains [23, 2]

$$\partial_t \hat{f}(t, \xi) = \int_{S^2} B \left( \frac{\xi \cdot n}{|\xi|} \right) \left[ \hat{f}(\xi^+) \hat{f}(\xi^-) - \hat{f}(\xi) \hat{f}(0) \right] dn \quad (7.6)$$

where  $\xi^{\pm} = \frac{\xi \pm |\xi|n}{2}$

**Theorem 5.**  $d_2(P_f, P_g) < d_2(f, g)$  for Maxwell molecules with cut-off collision kernel.

*Proof.* For Maxwell molecule with cut-off collision kernel  $\int B = S$ . Thus

$$\sup |Q^-| = \sup \left| \int B f_* d\Omega dv_* \right| = \sup |\rho S| < \infty.$$

Considering  $P = Q + \mu f = Q^+ + (\mu - Q^-) f$ , it is enough to prove  $d_2(Q_f^+, Q_g^+) < C d_2(f, g)$  for  $C$  big enough. Given

$$\hat{Q}_f^+ = \int_{S^2} B \left( \frac{\xi \cdot n}{|\xi|} \right) \left[ \hat{f}(\xi^+) \hat{f}(\xi^-) \right] dn,$$

one has

$$\frac{\hat{Q}_f^+ - \hat{Q}_g^+}{|\xi|^2} = \int_{S^2} B \left( \frac{\xi \cdot n}{|\xi|} \right) \left[ \frac{\hat{f}(\xi^+) \hat{f}(\xi^-) - \hat{g}(\xi^+) \hat{g}(\xi^-)}{|\xi|^2} \right] dn$$

From [28], one gets

$$\left| \frac{\hat{f}(\xi^+) \hat{f}(\xi^-) - \hat{g}(\xi^+) \hat{g}(\xi^-)}{|\xi|^2} \right| \leq \sup \left| \frac{\hat{f} - \hat{g}}{|\xi|^2} \right|$$

Thus, one has:

$$d_2(Q_f^+, Q_g^+) = \sup_{\xi} \left| \frac{\hat{Q}_f^+ - \hat{Q}_g^+}{|\xi|^2} \right| \leq S \sup \left| \frac{\hat{f} - \hat{g}}{|\xi|^2} \right| = S d_2(f, g)$$

□



## References

- [1] M. BENNOUNE, M. LEMOU, AND L. MIEUSSENS, *Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible NavierStokes asymptotics*, Journal of Computational Physics, 227 (2008), pp. 3781–3803.
- [2] A. V. BOBYLEV, *The Fourier transform method in the theory of the Boltzmann equation for Maxwellian molecules*, Akademiia Nauk SSSR Doklady, 225 (1975), pp. 1041–1044.
- [3] P. DEGOND, G. DIMARCO, AND L. PARESCHI, *The moment guided Monte Carlo method*, Int. J. Num. Meth. Fluids, 67 (2011), pp. 189–213.
- [4] P. DEGOND, S. JIN, AND L. MIEUSSENS, *A smooth transition model between kinetic and hydrodynamic equations*, J. of Comput. Phys., 209 (2005), pp. 665–694.
- [5] G. DIMARCO AND L. PARESCHI, *Fluid solver independent hybrid methods for multiscale kinetic equations*, SIAM J. Sci. Comput., 32 (2010), pp. 603–634.
- [6] G. DIMARCO AND L. PARESCHI, *Exponential Runge-Kutta methods for stiff kinetic equations*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 2057–2077.
- [7] G. DIMARCO AND L. PARESCHI, *Asymptotic preserving Implicit-Explicit Runge-Kutta methods for non linear kinetic equations*, arXiv:1205:0882, preprint (2012).
- [8] F. FILBET AND S. JIN, *A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources*, J. Comput. Phys., 229 (2010), pp. 7625–7648.
- [9] F. FILBET AND S. JIN, *An asymptotic preserving scheme for the ES-BGK model of the Boltzmann equation*, SIAM J. Sci. Comput., 46 (2011), pp. 204–224.
- [10] E. GABETTA, L. PARESCHI, AND G. TOSCANI, *Relaxation schemes for nonlinear kinetic equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2168–2194.
- [11] S. GOTTLIEB AND C.-W. SHU, *Total variation diminishing Runge-Kutta schemes*, Math. Comp, 67 (1998), pp. 73–85.
- [12] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, *Strong stability-preserving high-order time discretization methods*, SIAM Rev, 43 (2001), pp. 89–112.
- [13] E. HAIRER, S. P. N/ORSETT, G. WANNER, *Solving ordinary differential equations I. Non-stiff problems*, Springer Series in Comput. Mathematics, Vol. 8, Springer-Verlag 1987, 3rd edition (2008).
- [14] M. HOCHBRUCK, C. LUBICH, AND H. SELHOFER, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.
- [15] S. JIN, *Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review*, Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy), June 2010.

- [16] M. LEMOU, *Relaxed micromacro schemes for kinetic equations*, Comptes Rendus Mathématique, 348 (2010), pp. 455–460.
- [17] S. MASET AND M. ZENNARO, *Unconditional stability of explicit exponential Runge-Kutta methods for semi-linear ordinary differential equations*, Math. Comp., 78 (2009), pp. 957–967.
- [18] C. MOUHOT AND L. PARESCHI, *Fast algorithms for computing the Boltzmann collision operator*, Math. Comp., 75 (2006), pp. 1833–1852.
- [19] L. PARESCHI AND R. E. CAFLISCH, *An implicit Monte Carlo method for rarefied gas dynamics I: The space homogeneous case.*, Journal of Computational Physics, 154 (1999), pp. 90–116.
- [20] L. PARESCHI AND G. RUSSO, *Time relaxed Monte Carlo methods for the Boltzmann equation*, SIAM Journal on Scientific Computing, 23 (2001), pp. 1253–1273.
- [21] L. PARESCHI AND G. RUSSO, *Efficient asymptotic preserving deterministic methods for the Boltzmann equation*, AVT-194 RTO AVT/VKI, Models and Computational Methods for Rarefied Flows, Lecture Series held at the von Karman Institute, Rhode St. Genese, Belgium, 24–28 January (2011).
- [22] B. PERTHAME, *Boltzmann type schemes for gas dynamics and the entropy property*, SIAM Journal on Numerical Analysis, 27 (1990), pp. 1405–1421.
- [23] A. PULVIRENTI AND G. TOSCANI, *The theory of the nonlinear Boltzmann equation for Maxwell molecules in Fourier representation*, Annali di Matematica Pura ed Applicata, 171 (1996), pp. 181–204.
- [24] R. J. SPITERI AND S. J. RUUTH, *A new class of optimal high-order strong-stability preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.
- [25] C.-W. SHU, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, ICASE Report No. 97-65, (1997).
- [26] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii*, Journal of Computational Physics, 83 (1989), pp. 32–78.
- [27] S. TIWARI AND A. KLAR, *An adaptive domain decomposition procedure for Boltzmann and Euler equations*, Journal of Computational and Applied Mathematics, 90 (1998), pp. 223–237.
- [28] G. TOSCANI AND C. VILLANI, *Probability metrics and uniqueness of the solution to the Boltzmann equation for a Maxwell gas*, Journal of Statistical Physics, 94 (1999), pp. 619–637.
- [29] B. YAN AND S. JIN, *A successive penalty-based asymptotic-preserving scheme for kinetic equations*, preprint (2012).